

Motion Saliency Based Automatic Delineation of Glottis Contour in High-speed Digital Images

Xin Chen
Emerging Technology Center
Midea Group
San Jose, CA, 95134, USA
Email: chen1.xin@midea.com

Emma Marriott
Department of Computer Science
Stanford University
Stanford, CA 94305, USA
Email: emarriott@stanford.edu

Yuling Yan
Department of Department of Bioengineering
University of Santa Clara
Santa Clara, CA 95053, USA
Email: yyan1@scu.edu

Abstract—In recent years, high-speed videoendoscopy (HSV) has significantly aided the diagnosis of voice pathologies and furthered the understanding the voice production in recent years. As the first step of these studies, automatic segmentation of glottal images till presents a major challenge for this technique. In this paper, we propose an improved Saliency Network that automatically delineates the contour of the glottis from HSV image sequences. Our proposed additional saliency measure, Motion Saliency (MS), improves upon the original Saliency Network by using the velocities of defined edges. In our results and analysis, we demonstrate the effectiveness of our approach and discuss its potential applications for computer-aided assessment of voice pathologies and understanding voice production.

I. INTRODUCTION

Nowadays, high-speed videoendoscopy (HSV) has become more common in assessing voice condition and understanding mechanism of pronation [1][2][3]. In addition to reduce cost of HSV systems, this popularization can be attributed to two major developments. First, HSV systems are now capable of recoding speed at 2,000 - 4000 frames per second (fps)-a speed fast enough to resolve the actual vibrations of vocal folds. Second, the resolution of each frame has increased to 512×512 or higher, which allows each frame to preserve the real shape of subtle parts of the glottis contours.

In recent years, clinicians and speech scientists have developed numerous techniques to analyze HSV imaging data with a focus on the glottis [4][5][6][7]. In many of these techniques, automatic segmentation and delineation of the contours of glottis are both a critical and bottleneck step for successful analysis.

As shown in Figure 1, poor image quality and false or ambiguous edges complicate the task of segmentation of vocal folds from HSV images. Although some efforts have been made, there is not a perfect solution till now. Some of the proposed involve applying active contour model (Snake) [8], a well-known energy-minimization algorithm, to delineate the contour of glottis [9][10]. The common disadvantage of these methods is that the active contour model fails to delineate sharp concerns, such as top portion and bottom of the glottis. X. Chen et al. introduced a three-step region-growing method to segment glottal area [11]. However, the approach depended heavily on an accurate and reliable of a seed-pixel point. A. Mendez et. al employed Gabor filters to segment glottal

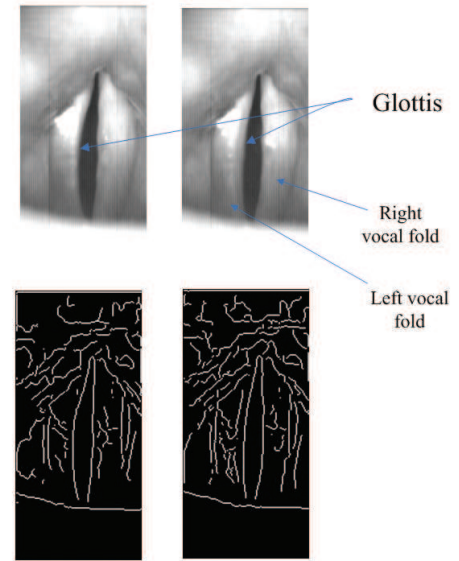


Fig. 1. Top: two adjacent frames of laryngeal images from an HSV imaging recording. The air space between the vocal folds, or the opening region, is glottis. The normal vocal fold motion in a sustained vowel phonation consists of open-close cycles. Bottom: Two edge images corresponding to top row images with Canny edge operator. As shown in top row, the image quality of the frames in HSV image is noisy. Most of the image frames don't contain rich texture information. In the bottom row, the edges are achieved by Canny edge operator. It is clear that contours of glottis in each image frame of HSV recording are salient curves.

area [12]. Although this approach is well-suited for images with rich texture information. HSV images frequently lack rich texture structures, as shown in Figure 1.

In this paper, we propose an improved Saliency Network to automatically delineate contour of glottis from HSV data. Our approach is based on original work on The Saliency Network introduced by Saashua and Ullman [13], which is considered as one of the top methods to solve the problem of extracting salient images with gap completion. Saliency measure of the Saliency Network, referred to as SU measure, is formulated from both local saliency and structural saliency. The SU

measure, however, is poor in computing structural saliency because it lacks enough properties [14], which sometimes leads to failure in finding real contours.

To address this issue, we propose a novel motion-derived saliency measure, called motion saliency (**MS**) measure. Since motion contains rich visual information on image structure and shape, our proposed MS saliency measure can effectively represent the missing structural saliency of the current Saliency Network.

This paper makes two major contributions as following:

- 1) We propose a novel MS measure that formulates from motion information in HSV imaging sequences. Combining with the MS measure, the Saliency network has greatly improved in structural saliency representation;
- 2) The improved Saliency Network has been successfully applied in automatical delineation of glottis contours from high-speed laryngeal images. These results have demonstrated its potential applications in the subsequent computer-aided assessment of voice pathologies and understanding voice production;

II. PROPOSED METHOD

A. Background of the Saliency Network

The Saliency Network aims at extracting salient contour from images and has been considered as one of the top bottom-up methods of perceptual group. In the Saliency Network, image pixels, also called elements [13], are classified into active and virtual elements. The elements that lie on edges are referred to as active elements, otherwise they are referred as to virtual elements.

Given a curve Γ composing of the $(N + 1)$ elements, the SU measure of Γ is defined as:

$$\Phi_{su}(\Gamma) = \sum_{j=i}^{i+N} C_{i,j} \rho_{i,j} \sigma_j \quad (1)$$

where σ_j is set to 1 for an active element, otherwise it is set to 0. The attenuation function $\rho_{i,j}$ is defined as:

$$\rho_{i,j} = \prod_{k=i+1}^j \rho_k \quad (2)$$

where $\rho_{i,i} = 1$, and ρ is in the range $(0, 1]$. If ρ_i is active it is set to as a value smaller than or equal to 1, and if ρ_i is virtual it is set as smaller than the value of the active. In [13], ρ is set to 1 if ρ_i is active, otherwise it is set to 0.7. $C_{i,j}$ aims to measure the shape of the curve that is inversely related to the total curvature of the curve and is defines as:

$$C_{i,j} = e^{-\int_{p_i}^{p_j} (\frac{d\theta}{ds})^2 ds} \quad (3)$$

where $\theta(s)$ is the slope along the curve s .

The SU measure defined in Equation 1 is a weighted contribution to the local saliency values σ_j along the curve. The weight consists of two factors, $C_{i,j}$ and $\rho_{i,j}$. The first inversely represents the number of virtual elements of the

curve, and the second inversely represents the total curvature of the curve.

The saliency of an element p_i is defined as the maximum of all curve starting from p_i :

$$\Phi(i) = \max_{\Gamma \in \Omega(i)} \Phi_{su}(\Gamma) \quad (4)$$

where $\Omega(i)$ denotes the set of curves starting from p_i .

B. Motion Saliency Measure

Since motion is an effective cue to represent image structure [15], MS measure calculated from motion cue effectively represents the structural saliency.

The formulation of the MS measure is based on different kinetic and kinematic properties corresponding to different types of motion. Since the image acquisition rate of HSV is some 10 times of the fundamental frequency of the vocal fold vibration, it is reasonable to be considered as a case of non-rigid object with low-order model shapes and slow variation load. As a matter of fact, most cases in medical images belong to this type of case of applications.

Given this type of motion pattern, the proposed MS measure is formulated based on the following two motion features:

- It has the maximum sum of the magnitude of velocity of all elements;
- The velocity of the adjacent elements of the contour are the closest. In other words, given any adjacent elements of the contour, both angular and spatial distances of the two vectors are closest;

Velocity Estimation The velocity is estimated from Lucas Kanade (**LK**) algorithm that is a widely used optical flow method [16]. While optical flow method is used to estimate rigid motion, the velocity values are reliable for non-rigid motion in some special applications, such as our work, for following two reasons:

- The motion of small patches between adjacent frames reasonably approximates to rigid motion.
- The observed brightness of any object point is constant over time;

As a result, the computation of MS measure based on these velocity estimations should also be reliable as well.

Normalization In order to smooth and normalize the velocity image estimated from LK algorithm, we take hyperbolic tangent function as a nonlinear transformation of each velocity image. The function is defines as:

$$f(x) = \frac{e^{\lambda_1 x} - e^{-\lambda_1 x}}{e^{\lambda_1 x} + e^{-\lambda_1 x}} \quad (5)$$

where λ_1 is a positive constant. In our work, we set λ_1 to 0.5.

Angular Distance Given the normalized velocity $\vec{V} = \{V_x(i), V_y(i)\}$ at the point i of the curve, the angular distance between point i and point j is the cosine of included angle of these two vectors, and is defined as:

$$D_a(i, j) = \frac{V_x(i) \times V_x(j) + V_y(i) \times V_y(j)}{\sqrt{V_x^2(i) + V_y^2(i)} \sqrt{V_x^2(j) + V_y^2(j)}} \quad (6)$$

where the value of D_a is in the range of $[-1, 1]$. A unity value of the angular distance implies that the two vectors are most-collinear or have maximum similarity, and -1 implies the opposite.

Spatial Distance Spatial distance aims to measure distance between two vectors in space, and is defined as

$$D_s(i, j) = \frac{2}{e^{-\lambda_2 D(i, j)} + e^{\lambda_2 D(i, j)}} \quad (7)$$

$$D(i, j) = \sqrt{(|V_x(i)| - |V_x(j)|)^2 + (|V_y(i)| - |V_y(j)|)^2} \quad (8)$$

where λ_2 is a positive constant. $D(i, j)$ is the Euclidean distance of two vector. The value of spatial distance is a positive constant in the range of $(0, 1]$. The λ_2 is set to 0.5 in our work.

MS Measure On the basis of the angular distance and the spatial distance, given a curve Γ composed of the $(N + 1)$ elements, we define the MS measure of Γ as:

$$\Phi_{ms}(\Gamma) = \sum_{j=i}^{i+N} M(i) D_a(i, j) D_s(i, j) \quad (9)$$

Where $M(i)$ is the magnitude of the normalized velocity $\vec{V}(i)$ at point i with

$$M(i) = \sqrt{V_x^2(i) + V_y^2(i)} \quad (10)$$

The MS measure defined in Equation 9 is a weighted contribution of $M(i)$ along the curve. The weight is a product of angle distance and spatial distance. If the two vectors are same, both the angular distance and the spatial distance are unity.

C. Proposed Improved Saliency Network

We define the improved Saliency Network as:

$$\Phi(\Gamma) = \alpha \Phi_{su}(\Gamma) + \beta \Phi_{ms}(\Gamma) \quad (11)$$

where α and β are weighted coefficient of SU measure and MS measure, respectively, and $\Phi(\Gamma)$ is a weighted sum of SU and MS measures.

Through the introduction of an additional term, $\Phi_{ms}(\Gamma)$, our method, compared to original Saliency Network $\Phi_{su}(\Gamma)$, improves on computation of structural saliency. The parameters α and β are set to the range of $[0, 1]$ and defined by users according for different cases.

Substituting $\Phi(\Gamma)$ for $\Phi_{su}(\Gamma)$ of the Equation 4, the saliency of an element p_i is rewritten as the maximum of all curve starting from p_i :

$$\Phi(i) = \max_{\Gamma \in \Omega(i)} \Phi(\Gamma) \quad (12)$$

where $\Omega(i)$ denotes the set of curves starting from p_i .

Essentially, the improved Saliency Network presents an optimization problem, and we adopted the dynamic programming is adopted to solve this problem.

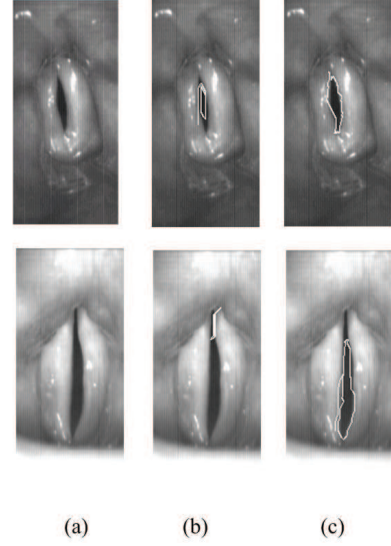


Fig. 2. (a): Frames of laryngeal images from an HSV recording. (b): Results of the Saliency Network. (c): Results from our proposed method.

III. RESULTS AND ANALYSIS

A. HSV Laryngeal Imaging

We used a Kay-Pentax (Lincoln, Park, NJ) HSV system to record the laryngeal images. This system is capable of acquiring images at a rate of 4,000 fps with a spatial resolution of 512×512 pixels.

During the examination, clinicians typically insert an endoscope into the subject's open mouth into the posterior pharynx at a 70-degree angle. To prevent gagging, clinicians took special care in avoiding the posterior pharyngeal wall and velum. The subjects are instructed to produce a sustained vowel /i/ while their tongues were secured. Phonation of the vowel /i/ retracts the epiglottis and base of the tongue, thus opening the laryngeal inlet and providing the best view of the larynx for imaging. The recording time for each HSV recording is typically within 2 seconds, yielding up to 8,000 image frames of data from this procedure.

B. Results and Analysis

Our experiment adopted a procedure of the subjective evaluation. The major criteria are to evaluate whether the results from glottis contour extraction are sufficiently accurate to facilitate further steps of processing with an ultimate goal of assisting in the computer-aided diagnosis of voice pathologies.

The parameters α and β of Equation 11 are set to 0.3 and 0.7, respectively. All of them are empirical values.

The results of analysis obtained from the proposed method as shown in Figure 2 demonstrated the robustness and accuracy of our method in extracting semantic contour from a noisy background. In comparison to Saliency Network, the computation of the MS measure efficiently represents saliency of structure. The effective motion saliency measure that we introduced enables our method to yield better results than the

Saliency Network. As clearly shown in Figure 2, our method outperforms the Saliency Network in extracting motion salient contour in noisy scenes from the HSV video.

IV. CONCLUSIONS

We proposed an improved Motion Saliency Network for automatic delineation of glottis contour from the HSV image sequences. The improved Saliency Network consists of two saliency measures, SU measure as defined in [13] and MS measure, which is a new motion-derived measure that we introduced in this paper. The MS measure is calculated from motion cues. Since motion efficiently represents global and shape properties of an image, the proposed model better resolves structural saliency properties and should outperform the Saliency Network.

We applied the proposed method to the analysis of sample HSV image sequences and the results have shown that our method was comparable to or better in performance than the existing state-of-the-art methods. Finally, although the improved Saliency Network is applied to glottis image segmentation herein, it is clear that this new model is a universal model because the formulation of the MS measure can vary from different kinetic properties of different types of motion. We anticipate that the proposed model should also be valid for extended applications in medical image analysis, provided reliable velocity estimation is possible.

ACKNOWLEDGMENT

The authors would like to thank Dr. Krzysztof Izdebski for his assistance in acquiring HSV data, some of which are used to test the validity of the proposed method.

REFERENCES

- [1] H. Larsson, S. Hertegard, P.A. Indestad, and B. Hammarberg, Vocal fold vibration: high-speed imaging, kymography and acoustic analysis, *Laryngoscope*, vol. 1000, pp. 2177-2122, 2000.
- [2] P. Mergell, H. Herzel, T. Wittenberg, M. Tigges, and U. Eysholdt, Phonation onset: vocal fold modeling and high-speed glottography, *The Journal of the Acoustical Society of America*, vol.104, no. 1, pp. 464-470, 1998.
- [3] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, U., and M. Döllinger, Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos, *Medical Image Analysis*, vol. 11, no. 4, pp 400-413, 2007.
- [4] Y. Yan, K. Ahmad, M. Kunduk, and D. Bless, Analysis of vocal-fold vibrations from high-speed laryngeal images Hilbert transform based methodology, *Journal of Voice*, vol. 19, no. 2 pp. 161-175, 2005.
- [5] J. G. Svec and H. K. Schutte, Videokymography: high-speed line scanning of vocal fold vibration, *Journal of Voice*, Vol. 10, No. 2, pp. 201-205, 1996
- [6] Y. Yan, X. Chen and D. Bless, Automatic tracing of vocal-fold motion from high-speed digital images, *Biomedical Engineering, IEEE Transactions on*, vol. 53 no. 7 pp. 1394-1400, 2006.
- [7] G. Chen, J. Kreiman, B. R. Gerratt, J. Neubauer, Y. L. Shue, and A. Alwan, Development of a glottal area index that integrates glottal gap size and open quotient, *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1656-1666, 2013.
- [8] M. Kass, A. Witkin, and D. Terzopoulos, Snakes: Active contour models, *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321-331, 1988.
- [9] D.D Mehta, D. D. Deliyiski, T. F. Quatieri, and R. E. Hillman, Automated measurement of vocal fold vibratory asymmetry from high-speed videolaryngoscopy recordings, *Journal of speech, language, and hearing research: JSLHR*, vol. 54, no. 1, 47, 2011.
- [10] Y. Yan, G. Du, C. Zhu, and G. Marriott, (2012, March). Snake based automatic tracing of vocal-fold motion from high-speed digital images, presented at the 2012 in Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on, pp. 593-596.
- [11] X. Chen, D. Bless, and Y. Yan, A segmentation scheme based on Rayleigh distribution model for extracting glottal waveform from high-speed laryngeal images, presented at IEEE-EMBS 2005, pp. 6269-6272.
- [12] A. Mendez, B. Garcia, I. Ruiz, and I. Iturricha, Glottal area segmentation without initialization using Gabor filters, presented at the 2008 in Signal Processing and Information Technology, IEEE International Symposium on, pp. 18-22.
- [13] S. Ullman, and A. Sha'ashua, Structural saliency: The detection of globally salient structures using a locally connected network, presented at 1988 International Conference on Computer Vision IEEE, pp. 321-327.
- [14] D. T. Alter, and R. Basri, Extracting salient curves from images: An analysis of the saliency network, *International Journal of Computer Vision*, vol. 27, no. 1, pp. 51-69, 1998.
- [15] G. Adiv, Determining three-dimensional motion and structure from optical flow generated by several moving objects, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 4, pp. 384-401, 1985.
- [16] B. D. Lucas, and T. Kanade, An iterative image registration technique with an application to stereo vision, *IJCAI*, vol. 81, pp. 674-679, 1981.